Abstract—Photonic artificial neural networks have garnered enormous attention due to their potential to perform multiply-accumulate (MAC) operations at much higher clock rates and consuming significantly lower power and chip real-estate compared to digital electronic alternatives. Herein, we present a comprehensive power consumption analysis of photonic neurons, taking into account global design parameters and concluding to analytical expressions for the neuron’s energy- and footprint efficiencies. We identify the optimal design-space and analyze the performance plateaus and their dependence on a range of physical parameters, highlighting the existence of an optimal data-rate for maximizing the energy efficiency. Following a survey of the best-in-class integrated photonic devices, including on-chip lasers, photodetectors, modulators and weighting elements, the mathematically calculated energy and footprint efficiencies are mapped into real photonic neuron deployment scenarios. We reveal that silicon photonics can compete with the best-performing currently available digital electronic neural network engines, reaching TMAC/s/mm² footprint- and sub-pJ/MAC energy efficiencies. Simultaneously, neuromorphic plasmonics, plasmo-photonics and sub-wavelength photonics hold the credentials for 1 to 3 orders of magnitude improvements even when the laser requirements and a reasonable waveguide pitch are accounted for, promising performance at a few fJ/MAC and up to a few pJ/MAC.

Index Terms—Energy Efficiency, Footprint Efficiency, Neuromorphic Photonics, Photonic Neuron, Power Consumption, Throughput

I. INTRODUCTION

Facing the data overflow challenge together with the lack of the adequate resources for its processing, that is commonly referred to as data deluge, ignited the interest in More-than-Moore and beyond-CMOS computing with the aim to overcome the limitations of von-Neumann computing architectures that form the stronghold of current general-purpose CPUs. The cost of transferring the data and/or instructions between the memory and CPUs becomes unbearable in terms of both latency and power consumption/dissipation [1], [2], [3], hindering the further increase in clock rates. This has forced the development of specialized computing architectures, relying on intertwining memory and CPU by moving from the CPU-centric to memory/data-centric computing [4]. These new architectures can be seen as the hardware counterpart of the already thoroughly studied artificial neural network (ANN) algorithms, to the extent that they are often jointly referred to as ANNs. More broadly, they can be classified as hardware accelerators, where ANNs comprise only a subcategory including also hardware for reservoir computing, decision making, information routing etc. [5] Accelerators are expected to be key-enablers of much-anticipated progress in cloud [6] and fog/edge computing [7], where time-sensitive applications with sub-ms latencies have to be supported, such as autonomous vehicles, augmented/virtual reality (AR/VR) etc.

At the same time, less than two decades from observing the 60-year valid trend of computation-per-joule doubling roughly every 1.5 years [8], Kooomy’s law found itself challenged as conventional computational machines started approaching the digital efficiency wall, set around 10 G-ops/J. This challenge brought important advances in neuromorphic computing, giving rise to computational machines such as IBM’s TrueNorth [9], [10], SpiNNaker [11], [12], [13], and Intel’s Loihi [14], [15], as well as hybrid analog-digital brain-inspired hardware as are Neurogrid [13], [16] and HiCANN [17], [18], [19]. Still, performing computation in electronic domain remains limited mostly to kHz regime, only with HiCANN reaching 22.4 MHz, yielding energy efficiencies of GMAC/s/W, with the exception of TrueNorth, raising to TMAC/s/W, and with corresponding footprint efficiencies usually in MMAC/s/mm² order of magnitude.

Building upon the well-known speed and energy benefits of photonic interconnects that are rapidly replacing electronics in information transmission at every hierarchy level of computing, the potential of photonics to yield high-performance neuromorphic infrastructures has already started to be researched in ANNs as well [6], [20], [21], [22]. Unlike the electronic counterparts, photonic ANNs detect and process information directly in optical domain, offering not only higher bandwidths, lower latencies and multiple options for signal multiplexing (wavelength, spatial modes, polarization), but also various nonlinearities not present, or difficult to attain, in electronics. Even though optical signal suffers from the noise associated with both the optical circuit and opto-electronic (OE) and electro-optic (EO) conversion [23], [24], precisions of 4 to 5-bits have been already shown to be

Manuscript received December 1, 2019. This work was supported by the EC through H2020 Projects PLASMONIAC (Contract 871391) and NEBULA (Contract 871658). (Corresponding author: Angelina Totović).

A. Totović, G. Dabos, N. Passalis, A. Tefas and N. Pleros are with the Department of Informatics, Aristotle University of Thessaloniki, 54621 Thessaloniki, Greece (email: angelina@auth.gr, ntamposg@csd.auth.gr, passalis@csd.auth.gr, tefas@csd.auth.gr, npleros@csd.auth.gr). 

Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap

Angelina Totović, George Dabos, Nikolaos Passalis, Anastasios Tefas and Nikos Pleros
feasible [23], [25] as well as sufficient, given the remarkable ANN error tolerance.

The main challenge for photonic ANNs is efficient vector-by-matrix multiplication (VMM), which is of great importance to signal and image processing, as well as NN training and inference [4]. Several photonic architectures emerged to answer this challenge, broadly classified as coherent and incoherent, with the three most prominent representatives being: (i) spatial light modulators (SLM), which can be made both as free-space and integrated [26], [27], (ii) the Broadcast-and-Weight (B&W) architecture, based on micro-ring resonator (MRR) banks and multiple wavelength channels [28], [29], [30], and (iii) coherent solutions, which either rely on cascaded coupled Mach-Zehnder interferometers (MZIs) for splitting/routing and weighting the signal [31], [32] or on parallel split-and-recombine dual-IQ modulator structures for input imprinting and signal weighing [33]. Regardless of the underlying architecture, the common goal of photonic ANNs is the simultaneous increase of throughput, energy- and footprint efficiency.

As indicated in the preliminary analysis of [25], neuromorphic photonics could be presumably assumed in the long run to yield energy efficiencies of sub-fJ/MAC with footprint efficiencies surpassing TMAC/s/mm² by 1-2 orders of magnitude. This has definitely shaped a forward-looking pathway for neuromorphic computing, placing photonics in the driver’s seat when taking into account mainly the fundamental physical limits enforced by the principles of light physics. However, this analysis is treating bandwidth and neuron fan-in as rather ample resources that are not rooted in current photonic technology capabilities, while at the same time the neuron footprint in [25] assumes off-chip lasers and pitch dimensions that don’t relate to currently feasible standards. To this end, it may presumably comprise a rather too optimistic conclusion, hard to associate with current and near-future photonic technology credentials.

In this paper, we adopt a holistic approach for computing energy and footprint efficiency calculation considering a generic but functionally complete single-layer neuromorphic photonic architecture where all photonic devices, including the laser, are included in the chip real-estate. By accounting for the photodetector’s sensitivity at the neuron’s output, as well as the power consumption of the modulator, we indirectly encompass the large part of the power consumption of the activation stage as well. Moreover, since the output of the PD can be used directly as the input modulation signal of the successive neuron layers, we expect no significant differences in terms of power consumption per neuron in multi-layer networks. We establish a theoretical framework that is subsequently used for analyzing the energy and footprint efficiency dependence on a range of important parameters including modulators, weights, neuron’s optical losses, photodiode, neuron fan-in and laser, identifying and explaining corresponding performance limitations and plateaus. Among the important findings is the dependence on optical data-rate, revealing that optical data-rate increases are not by default the panacea for energy efficiency improvements and designating an optimal bandwidth value for given operational parameters. We review state-of-the-art neuromorphic photonic architectures and relevant building blocks attempting to map our theoretical framework into technologically viable development scenarios, providing in this way technology roadmaps that comply with current and future photonic integration requirements. A new figure of merit relating energy and footprint efficiency is introduced, highlighting that neuromorphic plasmonics or plasmo-photonics and neuromorphic photonic crystal technologies can yield 1 to 3 orders of magnitude improved performance compared to top-performing electronic ANN engines.

This paper is organized as follows. In Section II we introduce a generic photonic N-to-1 neuron and systematize the metric parameters to be used for benchmarking different architectures on fair ground. In Section III we discuss the most prominent photonic neuron architectures available in the literature and map them to the generic N-to-1 neuron. In Section IV we study global design parameters and their influence on neuron performance. We infer the performance plateaus and discuss the design trade-offs in neuromorphic photonics. Section V offers an in-depth overview of the current state-of-the-art in photonic devices required for neuron actualization, as well as a technology roadmap, benchmarking different photonic platforms and implementation scenarios. In Section VI we draw the conclusions of the study.

II. ARTIFICIAL LINEAR NEURON

Figure 1(a) shows the abstract representation of an artificial McCulloch-Pitts neuron, a generalized form of perceptron [34], composed of axons carrying the input information $x_i$, synapses which weigh the inputs by multiplying them by $w_i$, an optional bias $b$, summation (or accumulation) stage ($\Sigma$) and an activation function triggering the neuron’s response, $\phi(\cdot)$, last two forming the somata of a neuron [35]. An artificial linear neuron, on the other hand, does not include the activation stage, and its output $y$ is recorded immediately following the summation stage, resulting in $y = \Sigma w_i x_i + b$. If linear network model is employed, $y$ can be directly forwarded to the following network layer, thus skipping the activation stage shown in Fig. 1(a). The number of inputs (equal to the number
of axons) defines the neuron fan-in, \( N \), an important metric parameter quantifying the network interconnectivity potential. Fan-in also affects the throughput, \( T \), broadly defined as the number of events per second, where by an event we assume multiply-accumulate (MAC) operation. In neuromorphic terminology, MAC operation computes the product of \( w_i \) and \( x_i \), and adds it to the accumulator \( y \), the output of a linear neuron. Aside from fan-in, the number of MACs per second is strongly influenced by data-rate \( B \), which is reflected in the definition \( T = NB \).

Fig. 1(b) maps the artificial linear neuron model to a generic photonic platform, allowing various photonic building blocks to be used for actualizing the constituent parts of the linear neuron from Fig. 1(a). The inputs \( x_i \) are usually realized through external modulation of the continuous wave (CW) carrier signal coming from a laser diode (LD) [29], [33], where the carriers’ wavelengths may differ between the branches (axons). The external modulators are required to support high data-rates, have low power consumption and low insertion loss (IL). Weighing mechanisms comprise the most diverse building block collection, ranging from MRRs [36] to ferroelectric memristors [37]. The accent in weight design is put on low power consumption in idle state, low IL and compactness. Broadly speaking, the most common mechanism exploited in weight design relies on attenuation-enforcing schemes, with the attenuation mechanism relying often on tunable filtering approaches, as shown in Fig. 1(c). The value of \( w_i \), pondering the signal’s intensity, can be varied between the limiting values (ideally 0 and 1) through repositioning the filter’s transfer function with respect to the signal’s wavelength. As shown in Fig. 1(c), this can be achieved either by sliding the filter’s transfer function through refractive index tailoring, which is a more common approach (e.g. through thermal control [38] or injecting the charge into the ring [39]) or by tuning the wavelength of the source [40], [41]. For implementation of negative weights (from -1 to 0), either an additional degree of freedom is required, such as phase or wavelength sign encoding, [33], [42], or a more elaborate architecture, relying on splitting the paths of positive and negative weighted signals [29].

Finally, depending on the neuron implementation scheme (single, double or multiple wavelengths) and the required domain of the output signal (optical or electrical), the accumulation stage in Fig. 1(b) may include multiplexer (MUX) followed by a photodiode (PD) [42], cascaded couplers [31], [33], a balanced photodetector (BPD) [29], etc. PDs are required to have high bandwidth, low sensitivity and a high responsivity, as well as low noise.

It should be stressed that, often, the functional boundaries between the constituent parts of an abstract neuron from Fig. 1(a) may not be as strict in the implementation given in Fig. 1(b) – a single photonic building block may play multiple roles, or a single function may require multiple photonic devices. Examples of multifunctionality could be weighing through the source wavelength tuning, or using a directly modulated laser (DML) both as a source and for input imprinting.

Assuming the \( \Delta h = 25 \mu m \) pitch between the axons, which complies with the high-density integration standards of current fabrication platforms, as well as \( L_{\text{wg}} = 5 \mu m \) length of the waveguides, the area occupied by a single linear photonic neuron, i.e., its footprint, can be estimated as:

\[
A = (N - 1)\Delta h \times L, \tag{1}
\]

where \( L \) stands for the total length of the neuron and according to Fig. 1(b) equals \( L = 4L_{\text{wg}} + L_{x} + L_{w} + L_{\Delta x} + L_{PD} \). As shown in Fig. 1(c), this can be achieved either by sliding the filter’s transfer function through refractive index tailoring, which typically ranges between 1% and 20% [40], [43], [44], [45], whereas \( P_{\text{X}} \) and \( P_{\text{W}} \) denote electrical powers consumed by a single modulator imprinting the input sequence \( x_i \), and a weighing mechanism for achieving the weight \( w_i \), be it by tuning the filter transfer function, or tuning of the lasing wavelength. In minimizing the power consumed, the optical input power ideally should not exceed the minimum required one, determined by the optical insertion losses in a light-path, \( \alpha \), given in dB, and the sensitivity of the photodetector \( P_{\text{R}} \), given in mW.

In the limiting case of the minimum optical power at the neuron input, total optical power reaching the PD must be equal to its sensitivity

\[
P_{\text{R}} = NP_{\text{S}}10^{-a[\text{dB}]/10}. \tag{3}
\]

After determining the minimum required optical power from (3), we rewrite (2) as

\[
P_{\text{el}} = P_{\text{R}}/\eta_{\text{wp}} \times 10^{a[\text{dB}]/10} + N(P_{\text{X}} + P_{\text{W}}). \tag{4}
\]

It should be noted that \( P_{\text{el}} \) implicitly depends on the data-rate \( B \), through the photodiode’s sensitivity. This dependence can be represented in generalized two-parameter form, either linear \( P_{\text{R}} \text{[mW]} = C_1(B/1 \text{GHz})^2 \), or logarithmic \( P_{\text{R}} \text{[dBm]} = C_1 + C_2 \log_{10}(B/1 \text{GHz}) \), where the two pairs of constants are related as \( C_1 = 10 \log10(c/1 \text{mW}) \) and \( C_2 = 10 c_2 \). Finally, (4) can be rewritten as

\[
P_{\text{el}} = C_1(B/1 \text{GHz})^2/\eta_{\text{wp}} \times 10^{a[\text{dB}]/10} + N(P_{\text{X}} + P_{\text{W}}). \tag{5}
\]

Although the footprint and the total power consumption are valuable metric parameters on their own, they cannot give an in-depth insight into the neuron performance irrespective of throughput. This is where footprint efficiency (FE), \( \eta_{\text{fe}} \), and energy efficiency (EE), \( \eta_{\text{ee}} \), step in, defined as ratios of throughput and footprint or electrical energy consumed, respectively.

\[
\eta_{\text{fe}} = \frac{NB}{(N - 1)\Delta h \times L}, \tag{6}
\]

and

\[
\eta_{\text{ee}} = \frac{NP_{\text{S}}10^{-a[\text{dB}]/10}}{NP_{\text{S}}10^{-a[\text{dB}]/10}} + N(P_{\text{X}} + P_{\text{W}}). \tag{7}
\]
\[\eta_E = \frac{NB}{c_1(B/1 \text{ GHz})^2/\eta_{wp} \times 10^{a[\text{dB}]}/10 + N(P_X + P_W)} . \quad (7)\]

Unlike throughput, which can be increased either by increasing \(N\) or \(B\) or both, same does not hold for the quantities defined by (6) and (7). In the limiting case for large fan-in, \(\eta_e\) approaches to \(B/(\Delta h L)\), implying that it can be increased either by increasing the data-rate, or by reducing the pitch of axons or the length of a neuron in other words, the average area per axon. Similarly, in the limiting case for \(N \to \infty\), the value of \(\eta_e\) approaches to \(B/P_x + P_w\), implying that the most significant contribution to EE from total energy consumption comes from modulators and weighting mechanism and that the increase in \(\eta_E\) can be ensured by increasing the data-rate. However, this conclusion does not hold for moderate fan-in values anticipated in photonic configurations, where the power consumed by optical sources dominates the \(P_{el}\), especially in the case of low power consumption by \(x_i\) and \(w_j\). Moreover, if the coefficient \(c_2\) is greater than 1, an optimal data-rate for a maximum energy efficiency can be determined

\[B_{opt} = \frac{N(P_X + P_W)}{c_1(c_2 - 1)} \eta_{wp} 10^{-a[\text{dB}]}/10^{1/c_2} \times 1 \text{ GHz} . \quad (8)\]

Inserting (8) into (7) gives the plateau value of \(\eta_e\) for \(c_2 > 1\)

\[\eta_{E, max} = \left(1 - \frac{1}{c_2}\right) \frac{1 \text{ GHz}}{P_X + P_W} \times \left[\frac{N(P_X + P_W)}{c_1(c_2 - 1)} \eta_{wp} 10^{-a[\text{dB}]}/10^{1/c_2}\right] , \quad (9)\]

and the corresponding optimal power consumption

\[P_{el, opt} = N(P_X + P_W)\left[(c_2 - 1)^{-1} + 1\right] . \quad (10)\]

Finally, an alternative figure-of-merit to EE is energy per MAC event, \(E_{MAC}\), which is an inverse of \(\eta_E\), and is given in units of J/MAC.

III. PHOTONIC NEURON ARCHITECTURES

Before proceeding with the analysis of the photonic neuron performance according to the metric parameters outlined in Section II, we briefly describe the most prominent representatives of noncoherent and coherent classes, given in Fig. 2, and map them to the generic photonic linear neuron represented in Fig. 1(b). It should be stressed that the architecture shown in Fig. 1(b) is \(N\)-to-1, whereas some of the photonic architectures in the available literature are designed having the full-connectivity in mind. As such, they achieve their maximum performance in \(N\)-to-\(M\) configurations, where the number of outputs \(M\) is often equal to the fan-in \(N\). In order to compare the architectures on fair ground, we identify the minimum number of components required for generating a single output from the \(N\)-to-\(M\) architectures, i.e., assuming \(M = 1\), thus mapping them to \(N\)-to-1 model.

Broadcast-and-weight architecture, shown in Fig. 2(a), has been proposed in [28], and has been elaborated in more detail in [29], [30], [36]. It can be used in both feed-forward and recurrent configuration, depending on the routing of the outputs. It relies on multiple-wavelength operation, with each input being assigned a unique wavelength (\(\lambda_i\)), making the architecture non-coherent. Layout given in Fig. 2(a) shows a fully connected recurrent configuration, where the output generated in one cycle, \(y_j = \Sigma w_{ji} x_i\), with \(j = 1..M\) and \(i = 1..N\), is used as the input for the next cycle, \(x_i\), which requires the identical number of inputs and outputs, \(M = N\). When used in feed-forward configuration, the number of outputs \(M\) may differ from fan-in \(N\). The input data \(x_i\) is imprinted externally onto a CW optical carrier emitted from a designated LD via Mach-Zehnder modulator (MZM), as shown in Fig. 2(a), which corresponds to the generic neuron depicted in Fig. 1(b). The inputs are then multiplexed into a single waveguide using arrayed waveguide grating (AWG), and further split into \(M = N\) branches, each of which will serve as an input to one of the \(M\) parallel \(N\)-to-1 neurons. The waveguide carrying the \(N\) wavelength-multiplexed signals is routed next to the MRR weight bank constituted of \(N\) MRRs, each with the weight \(w_{ji}^*\), ideally varying between 0 and 1, where \(j \in \{1,..,M\}\) is the index of the parallel neuron and \(i \in \{1,..,N\}\) denotes the input being weighted. Each MRR filter is designed such that its transfer function can be continuously tuned, as shown in Fig. 1(c), achieving controlled attenuation of the signal’s power at the corresponding \(\lambda_i\). Depending on the relative
position of \( \lambda \), with respect to the filter’s TF, a fraction of the power, proportional to \( w_{j,i}^* \), will be sent to the "through" port (THRU), continuing to travel along the waveguide, whereas the remaining fraction, proportional to \( w_{j,i} \), will be sent to the "drop" port (DROP), where it will be collected by another, parallel waveguide. The resulting signal will be proportional to the subtraction result between the THRU and DROP powers, 
\[
y_j = \Sigma (w_{j,i}^* - w_{j,i}) x_i = \Sigma w_{j,i} x_i,
\]
ideally yielding 
\[
y_j = \Sigma (2w_{j,i}^* - 1) x_i \quad \text{for} \quad w_{j,i} = 1 - w_{j,i}^*,
\]
implies that the resulting weight \( w_{j,i} \) can take negative values as well depending on the ratio of the powers sent through the two MRR ports. Comparing Fig. 2(a) to Fig. 1(b), each channel within the waveguide of B&W architecture corresponds to one axon, and no MUX is required at accumulation stage as channels continue to travel together down the THRU or DROP waveguide. It is important to notice that MUX alone does not play the role of an accumulator, regardless of its position before or after weighing, as the channels are still separable within each waveguide, as well as between waveguides. The accumulation is done during detection by the two photodiodes placed in a balanced configuration, as shown in Fig. 2(a), such that the electrical signal resulting from DROP port is being subtracted from the electrical signal resulting from THRU port, yielding the neuron output.

Fully connected \( N \)-to-\( M \) neuron based on B&W architecture requires \( N \) lasers, \( N \) modulators, \( N \times M \) weights and \( M \) balanced photodetectors, which is a minimum for noncoherent architecture. The maximum performance in \( N \)-to-\( 1 \) configuration, i.e., for \( M = 1 \), would be achieved with the same number of lasers and modulators, \( N \) weights and 1 balanced photodetector. However, omitting parallelization removes the need to split the multiplexed inputs into \( M \) branches, thus reducing the required optical power leaving the laser to compensate for the same optical losses in a single light-path. Therefore, the electrical power consumption of lasers, in ideal case of exactly compensating the losses, can be approximately \( M \) times lower in \( N \)-to-\( 1 \) configuration comparing to the fully connected \( N \)-to-\( M \) case.

Wavelength-encoded-input architecture offers a convenient possibility of implementing the weights \( w_{i} \) not as discrete components, but rather through sources’ wavelength tuning, as illustrated in Fig. 1(c). A filtering function is still required along the optical path; however, it can be made passive, and without the need to bias it. Moreover, filtering property of the multiplexer, such as AWG, can be exploited. Although certain energy will still be required for laser wavelength tuning, which can be done by thermal or any other suitable effect, additional energy for biasing of the weight as a discrete component is not required, and the area occupied by a neuron is reduced, suggesting the potential in increasing of both EE and FE.

Another category of neurons operates coherently, performing algebraic operations on the electrical field amplitude rather than the signal power and harnesses the interference as its accumulation mechanism. An example from this category is the recently developed coherent optical linear neuron (COLN), shown in Fig. 2(b), which exploits a dual-IQ modulator as its basic algebraic cell and supports both positive and negative weights by exploiting the signal’s phase, using just a single wavelength [33]. Other suggestions for coherent operation that can be found in literature rely on wavelength diversity for sign notation [42], and require twice as much electrical energy for biasing of lasers, while halving the WDM parallelization potential of COLN. Neither of coherent architectures requires full connectivity to reach its maximum performance. In fact, both are easily mapped to \( N \)-to-1 model, and at the same time support seamless scaling up to \( N \)-to-\( M \) architectures by exploiting different wavelengths for carrying different input sequences.

Instead of using \( N \) different laser sources for the CW seed signals as shown in Fig. 1(b), COLN requires exactly one LD, whose signal is split to \( N+1 \) branches, Fig. 2(b), preserving the coherence. Of these, \( N \) are used as axon seeds within the core of the linear neuron, i.e., the optical linear algebra unit (OLAU), while the remaining one is used as a bias for converting the sign information of the accumulated signal from phase to intensity domain. In this manner, the number of lasers is reduced \( N \) times, although the required optical power to overcome the losses in the optical path is increased approximately \( N \) times. Input signals \( x_i \) are imprinted by external modulation using MZMs, as shown in Fig. 2(b), and corresponding to Fig. 1(b). COLN architecture requires two distinct modulators for weighting, one for electrical field amplitude attenuation, proportional to \( |w_i| \), with the underling mechanism shown in Fig. 1(c), whereas the other serves for phase modulation, where \( 0/\pi \) denotes positive/negative weight, or \( \text{sgn}(w_i) \). For proof-of-principle, in [33], dual-IQ modulator was used as an elementary two-input OLAU, as depicted in Fig. 2(b), implying that MZM was used for signal weighting. However, relaxed conditions regarding weight update speed allow for using of other photonic components which introduce lower IL, occupy less area and/or require no biasing. Finally, the signals leaving each axon, \( w_i/\text{sgn}(w_i)x_i/N \), are coherently recombined within the accumulation stage, constituted from cascaded Y-junction combiners, yielding \( \Sigma (w_ix_i)/N \). This signal is brought to interfere with the properly attenuated and phase modulated signal from the bias branch, \( w_b\exp(j\phi_b) \), acting as a reference level, Fig. 2(b), giving \( w_i\exp(j\phi_b + \Sigma (w_ix_i)/N) \), thus converting the sign of the sum from phase to intensity domain. This additional bias interference step is an integral part of the accumulation stage. Finally, the signal is forwarded to the photodiode, which serves only for OE conversion, and is not required in accumulation stage. This implies that signal leaving COLN is compatible both with optical and electrical activation functions, if one is desired.

A coherent neuron architecture, envisioned as a core of deep learning (DL) photonic ANNs for providing fully optical VMM, is presented in [31]. It is primarily designed as \( N \)-to-\( M \) architecture, although it can support \( M = 1 \) version. The core of the neuron is the optical interference unit (OIU) realized through cascaded MZIs arranged as per Fig. 2(c), following the layout proposed by Clement et al. [46], which in turn comprises an improvement of the layout proposed by Reck et
Implementing VMM relies on singular value decomposition (SVD) of the real valued matrix $M$ to $M = U \Sigma V^T$, requiring only beam splitters and phase shifters for unitary transformations $U$ and $V$. as shown in Fig. 2(c), (d). On the other hand, the rectangular diagonal matrix $\Sigma$ is implemented by using attenuators or amplifiers, which can again be implemented through cascaded MZIs, as shown in [31]. The imprint of the electrical input signal $x_i$ onto an optical input signal, shown in Fig. 1(b), is not envisioned as a part of the neuron architecture and it is done externally. As the architecture is coherent, only a single wavelength source is required, with the constrain of preserved coherence and enough optical power to allow splitting to $N$ axons and overcoming the losses along the optical path.

The unitary matrix $U$ can be seen as routing/mixing stage preceding the weighing, not present in the model from Fig. 1(b), whereas $V^T$ corresponds to the interference accumulation stage following the weighing, as shown in Fig. 1(b). The weights $w_i$ require a minimum of $\min\{N,M\}$ attenuators, implemented by MZIs in [31]. According to [46], [47], controlling of the splitting ratio in Fig. 2(c) (and thus the routing/mixing) and accumulation is done by employing a maximum of $N(N-1)/2$, i.e., $M(M-1)/2$ of variable beam splitters of Fig. 2(d) per unitary matrix, respectively. The MZIs are programmed relying on internal ($\theta$) and external ($\phi$) thermo-optic phase shifters. Such an elaborate input rerouting and accumulation stage puts a burden both on energy consumption (through internal and external phase control) and footprint due to $[N(N-1) + M(M-1)]/2 + \min\{N,M\}$ of MZIs in total, especially if weight update is anticipated, which also requires the update of the external/internal phases of MZIs from Fig. 2(d) within unitary matrices in Fig. 2(c). For a single output, the number of MZIs reduces to $[N(N-1)+1]/2$, still scaling with fan-in following the square law, unlike the B&W or COLN architectures which scale linearly with $N$.

In conclusion, the complexity and uniqueness of the architecture from [31], relying on multiple paths the signal can take, does not allow for it to be mapped to Fig. 1(b) model with one-to-one correspondence. As the metric parameters derived in Section II correspond to model of Fig. 1, we omit OIU from [31] and Fig. 2(c), (d) in the detailed performance analysis that follows.

IV. ENERGY EFFICIENCY ANALYSIS AND DISCUSSION

In order to understand the potential of photonics in ANNs, we set to investigate neuron power consumption and energy efficiency for a broad range of architectures, by varying the values of the fan-in, $N \in [2^1,2^7]$, the insertion loss along the optical path, $\alpha \in [0.5, 30]$ dB, the combined electrical power consumed by the modulator and the weight within a single axon, $P_x + P_w \in [0.1, 50]$ mW, and the receiver sensitivity, $P_R \in [-25,20]$ dBm. In all of the analyzed cases, we choose the laser wall-plug efficiency of $\eta_{wp} = 10\%$.

Employing (5) and (7), derived in Section II, requires correlation of the receiver sensitivity with the data-rate.
In the case when both \(\alpha\) and \(P_R\) are independent of \(N\), as shown by the dashed lines in Fig. 4(c). Dashed black line bounding the shaded region shows the lower limit of the neuron power consumption, limited by first term of (4). Dashed black lines bounding the shaded regions show the lower limit of the neuron power consumption, limited by second term of (4). (PD: photodiode)

1.28\times10^3\) mW in Fig. 4(c). In the case when both \(\alpha\) and \(P_R\) are low, total power consumption is bounded by the second term of (4), which scales linearly with \(N\) and \(P_X + P_W\) and is independent on \(\alpha\) and \(P_R\). Once the losses and/or sensitivity are increased, first term of (4) reclaims the dominance and \(P_{el}\) becomes independent of \(N\) and/or \(P_X + P_W\), approaching to 10\(^6\) mW per neuron for \((\alpha, P_R) = (30\ dB, 20\ dBm)\), the same value for all three analyzed cases from Fig. 4. In should be noted that such high values of \(\alpha\) and/or \(P_R\) are not anticipated in real deployed systems, and are used only for indicative purpose.

The influence of fan-in and modulation power consumption on \(P_{el}\) is examined in more detail in Fig. 5(a), for a fixed pair of losses along the optical path and PD sensitivity of \((\alpha, P_R) = (17\ dB, -14\ dBm)\), corresponding to the flat area of Fig. 4(c). According to (4), the power consumption of the source(s) in this case is approximately 20 mW, denoted by the dashed line bounding the shaded region in Fig. 5(a). This value corresponds to \(P_{el}\) for negligible \(P_X + P_W\), i.e., the origin of the collection of curves given in Fig. 5(a). As \(P_X + P_W\) increases, so does the \(P_{el}\), following the linear law, and same can be seen for the increase in \(N\), which directly increases the number of active components, and therefore the neuron power consumption.

Figures 5(b)-(c) show 2D cross-section of Figs. 4(a)-(b) for different PD sensitivities. The lower limit of \(P_{el}\) is determined by the second term of (4) and yields 12.8 mW for Fig. 5(b) and 128 mW for Fig. 5(c), as shown by the dashed lines bounding the shaded regions in respective figures. As the first term of (4) becomes dominant with the increase of \(\alpha\) and/or \(P_R\), the dependence of \(P_{el}\) on \(\alpha\), as well as \(P_R\), becomes linear. We once again confirm that for low to moderate fan-ins, power consumption of the laser, needed to overcome the IL and meet PD sensitivity requirements, usually dominates \(P_{el}\).

### B. Energy Efficiency

Power consumption alone offers very coarse guidelines in optimizing the linear neuron, suggesting reduction in fan-in, modulation power consumption and losses along the optical path, as well as improvement of PD sensitivity. However, the computational yield, given by throughput \(T\), imposes an independent and conflicting set of criteria, requiring high fan-ins and data-rates. Therefore, we proceed with the analysis of energy efficiency \(\eta_E\), defined by (7), and its inverse, energy per MAC \(E_{MAC}\), taking into account both \(P_{el}\) and \(T\).

Figure 6 shows the 3D dependencies of energy efficiency \(\eta_E\) [Figs. 6(a)-(c)], and energy per MAC \(E_{MAC}\) [Figs. 6(d)-(f)], on data-rate, \(B\), and losses along the optical path, \(\alpha\), for a fan-in of \(N=128\) and three values of the modulation power consumption, \(P_X + P_W = \{0.1, 1, 10\}\) mW. Looking at all three analyzed cases, similarities in \(\eta_E\) and \(E_{MAC}\) dependencies on \(B\) and \(\alpha\) can be observed for various modulation powers. The most prominent feature is the occurrence of maximum in \(\eta_E\) as \(\alpha\) increases, Figs. 6(a)-(c), and the corresponding minimum in \(E_{MAC}\) as \(\alpha\) increases, Figs. 6(d)-(f). The existence of extrema has already been predicted in Section II, through (9) for an optimal data-rate given by (8), conditioned by \(c_1 > 1\), which indeed is the case from Fig. 3. As long as IL is fairly low and \(P_X + P_W\) high enough, \(B_{opt}\) predicted by (8) will be on the limit...
of the data-rates attainable in practice or beyond them, e.g., higher than 100 Gb/s for \( P_X + P_W = 10 \text{ mW} \), Fig. 6(c), allowing for the improvement of \( \eta_E \) with the increase of \( B \) for all cases of practical interest. As \( P_X + P_W \) decreases and/or \( \alpha \) increases, the optimal data-rate moves towards lower values, dropping to 10 Gb/s or even Gb/s-range and the plateau of \( \eta_E \) is being reached for lower \( B \). Increasing the data-rate beyond this point will continue to increase the throughput \( T \), however, the significant increase in power consumption will bring \( \eta_E \) down. Estimated values of \( \eta_E^{\text{max}} \) from Figs. 6(a)-(c) are in the range of \( 10^4 \) to \( 10^5 \) GMAC/s/W, depending on \( P_X + P_W \) and \( \alpha \), corresponding to sub-pJ energies per MAC of 0.1 to 0.01 pJ/MAC, respectively, shown in Figs. 6(d)-(f).

Increasing the modulation power consumption from Figs. 6(a), (d) to Figs. 6(c), (f), reveals either an overall decrease in efficiency for lower data-rates and losses, or induces no visible change for high \( B \) and \( \alpha \). As long as \( B \) and \( \alpha \) are low, power consumption is governed by the second term of (4), leading to linear increase of \( P_{el} \) with \( P_X + P_W \), which manifests as a drop in \( \eta_E \) (increase in \( E_{MAC} \)). Once \( B \) and/or \( \alpha \) are high enough for the first term of (4) to dominate power consumption, \( P_{el} \) becomes less sensitive to variation of \( P_X + P_W \), which implies that no significant changes will appear in \( \eta_E \) and \( E_{MAC} \) from Figs. 6(a), (d) to Figs. 6(c), (f).

Since fan-in affects both \( T \) and \( P_{el} \), we explore in more detail the dependence of \( \eta_E \) on \( N \) in Fig. 7(a) for a range of \( P_X + P_W \) and a fixed pair of \( (\alpha, B) = (17 \, \text{dB}, 18 \, \text{Gb/s}) \), yielding PD sensitivity of -14 dBm, corresponding to the flat area of Fig. 4(c). We observe that \( N \) influences \( \eta_E \) (as well as \( E_{MAC} \)) only for very low \( P_X + P_W \), up to approximately 10 mW, for a given pair of \( (\alpha, B) \). In this case, the first term of (5) yields approximately 20 mW of source’s power consumption, whereas the second term of (5) dominates, being roughly \( N \) times greater. Having in mind that both \( T \) and the second term of (5) scale linearly with \( N \), while the contribution of the first term of (5) is constant, \( N \) will have a negligible influence on \( \eta_E \) as long as the second term of (5) dominates power.
consumption. Therefore, \( \eta_E \) will approach its limiting value of \( B(P_X + P_W) \) as discussed in Section II, yielding 1.8x10³ GMACs/s/W and 0.55 pJ/MAC for modulation power consumption of 10 mW.

2D cross-section of Figs. (a)-(b) for \( N = 128 \) and different \( B \), shown in Figs. (b)-(c), offers a clearer perspective on data-rate influence on \( \eta_E \) for a range of \( \alpha \), depicting also the plateau \( \eta_E^{\text{max}} \), determined for \( B_{\text{opt}} \). Decreasing the IL increases \( \eta_E \) linearly, as long as the data-rate is beyond the \( B_{\text{opt}} \), implying that power consumption of the source(s) dominates the total power consumption. For lower data-rates, IL shows negligible influence on \( \eta_E \), showing the dominance of \( P_X + P_W \) to power consumption. Finally, Fig. 7(c) shows that increase in data-rate has beneficial influence on \( \eta_E \) for moderate \( P_X + P_W \) and low losses, as long as \( B < B_{\text{opt}} \). As the existence of optimal data-rate has important implications to energy efficiency, we proceed with a thorough analysis of \( B_{\text{opt}} \) and \( \eta_E^{\text{max}} \).

C. Optimal Data-Rate and Energy Efficiency Plateau

In the analysis of the optimal data-rate and EE plateau we rely on (8)-(10), derived in Section II, as well as the ranges of parameter values outlined in Section IV. Figure 8 shows the dependence of optimal data-rate \( B_{\text{opt}} \) [Figs. 8(a)-(c)] and the corresponding EE plateau \( \eta_E^{\text{max}} \) [Figs. 8(d)-(f)] on loss along the optical path \( \alpha \) and cumulative modulation power consumption \( P_X + P_W \) for various fan-ins \( N \), namely 8, 32 and 128. Optimal data-rate can be interpreted as the data-rate that provides high enough throughput \( T \) while still keeping \( P_R \) low, therefore reducing the electrical power consumption of the source(s) required to provide sufficient optical power to meet the PD sensitivity criterion. In other words, \( B_{\text{opt}} \) brings the power consumption originating from source(s) to comparable level with the power consumption originating from the modulators and weights, as shown by (10). As \( \alpha \) is reduced, higher \( P_R \) is allowed for the same power consumption of the source(s), and therefore \( B_{\text{opt}} \) increases, as shown in Figs. (a)-(c) and confirmed by (8). Moreover, decrease in \( \alpha \) brings the EE plateau to higher values since it increases the numerator of (7) through increase of \( B_{\text{opt}} \), whereas the denominator remains defined by (10), as confirmed in Figs. (d)-(f) and (9). Similarly, increase in \( P_X + P_W \) strengthens the contribution of the second term of (5) to \( P_d \), requiring higher data-rates so that power consumption originating from the source(s) becomes comparable with the modulation one. As shown in Figs. (a)-(c) and (8), this also results in \( B_{\text{opt}} \) increase. However, a reverse effect of \( P_X + P_W \) on \( \eta_E^{\text{max}} \) can be observed in Figs. (d)-(f) and (9). This trend is expected since the total electrical power consumption, given by (10), is increased with a higher slope than the throughput, proportional to (8). Finally, following the increase of \( N \) from Figs. (a), (d) to Figs. (c), (f), we observe an increase in both \( B_{\text{opt}} \) and \( \eta_E^{\text{max}} \), supported by (8) and (9). The underlying cause is related to lower sensitivity of \( P_d \) on energy consumption by source(s), allowing for increase in first term of (5) in order to become comparable with the second term of (5). At the same time, looking at the rewritten (9), based on (7) and (10),

\[
\eta_E^{\text{max}} = \frac{B_{\text{opt}}}{(P_X + P_W)(c_2 - 1)^{-1} + 1},
\]

we reveal that \( \eta_E^{\text{max}} \) scales directly with \( B_{\text{opt}} \), and therefore, indirectly with \( N^2/(1/c_2) \).

V. STATE-OF-THE-ART PHOTONIC DEVICES AND PHOTONIC NEURON TECHNOLOGY PERSPECTIVES

After examining the performance characteristics of a linear neuron using a more global parameter search space, we proceed with a survey of photonic devices that could be
potentially exploited for realizing all functional building blocks required by a photonic neuron. More specifically, performance metrics for best-in-class devices were tabulated including photodetectors (Table I), modulators (Table II), weighting elements (Table III) and lasers (Table IV). The main purpose of this survey has been to classify enabling technologies and map potential photonic neuron deployments onto current and near-future technology capabilities, concluding to energy- and footprint efficiencies that could be considered within a pragmatic reach of today’s technology perspectives. Treating laser and photodiode resources as well as neuron fan-in parameters under the prism of a technologically viable perspective, our analysis can elucidate on possible neuromorphic technology platforms over a realistic development roadmap, filling the gap between available digital electronic and projected photonic ANN technology, where energy consumption and real-estate parameters are dictated solely by physical law boundaries.

Before presenting in detail our deliberate choices regarding the potential photonic neuron technology roadmaps, we proceed with a brief overview on advances and progress witnessed for state-of-the photonic devices.

A. Photodetectors

Starting from photodetectors, a neuron should ideally accommodate PD devices with high bit-rates, low power consumption and low sensitivity, a selection of which is given in Table I. PDs operating at the photovoltaic mode, hence being bias-free, become alluring towards minimizing the overall power consumption of the neuron. Such photodetectors have been reported for photonic crystal (PhC) and silicon-germanium (SiGe) technologies supporting bit-rates of up to 40 Gb/s [57], [58], [59], [60]. Seeking for PDs with higher bit-rates, SiGe technology [52], [53], [59], [60] steps into, where bit-rates climbed up to 100 Gb/s for a PIN based SiGe PDs, yet being operated at the photo-conductive mode [52]. Promising alternatives resorting on plasmonics have been also very recently reported, still requiring of course further developments to optimize performance by minimizing interface losses and bias voltages [51], [55], [56]. Apart from high-speed operation and low-power consumption, sensitivity plays also an important role dictating the loss budget that a linear neuron can tolerate. In this realm, PhC nanophotonic PDs form an almost unbeatable choice by being capable to detect optical signals with 1 μW optical power, negating the need for amplifiers. Despite the progress made in the field of CMOS transimpedance amplifiers (TIAs) (<100 fJ/bit), their utilization should be ideally avoided in order to maintain circuits’ footprint as well as complexity low.

B. Modulators

Low-loss, energy-efficient and high-speed modulators (Table II) are also imperatively needed to allow photonic neurons to process data with speeds higher than that of digital electronics. A common requirement for all types of modulators for keeping energy consumption at the lowest possible envelope is to exhibit low insertion losses and operate under CMOS driving voltages, i.e., ~1 Vpp, so as to avoid the need for RF driver amplification stages that in most cases bring power consumption to a few pJ/bit. During the last decades, rigorous developments have led to a large variety of MZI- and resonator-based modulators. Among different
technologies, plasmonics revolutionized conventional photonic modulators by dramatically reducing their dimensions, while offering at the same time optical modulation speeds to 120 Gb/s with attojoule-per-bit power consumption [63]. Compact plasmonic assisted ring resonator-based modulators have been also reported for 72 Gb/s Non-Return-to-Zero (NRZ) modulation formats [64]. While plasmo-photonic modulators are still taking up the gauntlet to reduce optical propagation losses and permeate CMOS manufacturing in all tiers, alternative solutions can be found in silicon-based counterparts. In this context, electro absorption modulators (EAM) based on Ge and III-V PhCs on Si, as well as Si MRR substitutes, have been reported providing modulation rates up to 56 Gb/s. Modulators using more exotic materials such as Indium Titanium Oxide (ITO) have been also demonstrated with the prospect to reach atto-Joule consumption and ultra-compact layouts [68], [69].

C. Weighting Elements

Carrying out on-chip weighting requires in most cases tunable filter devices with ultra-low power consumption in order to perform inference tasks competitively against electronics. An overview of some of the weight candidates is given in Table III. Exploiting silicon photonic devices with thermo-optic phase shifters lead to power consumption levels from a few to tens mW, which is prohibitively high [38]. Alternative solutions incorporating ferroelectric and III-V materials have emerged to realize energy efficient phase shifters with power consumption in the nW regime [73], [74], [75]. Moreover, post-fabrication trimming techniques were demonstrated towards zeroing power consumption at all in interferometric and resonant devices, yet lacking of reconfigurability [76]. Requiring an extra optical carrier, all-optical phase shifters using phase change materials have been also proposed [77].

D. Lasers

Energy-efficient on-chip laser sources are always of upmost importance, since they are expected to dominate the total energy consumption of a photonic neuron for reasonable fan-in values. Performance metrics have been summarized in Table IV for various lasers that can be integrated on Si waveguide platforms, either through heterogenous or hybrid integration. The selection of a certain laser for being considered as the optical power supply in a specific neuron layout would be predominantly determined by the available power loss budget and the sensitivity of the employed PD.

E. Photonic Neuron Deployment Scenarios and Technology Roadmaps

Based on the silicon-compatible state-of-the-art devices identified for all necessary functional blocks, three possible technology vehicles for realizing photonic neurons can be easily defined along the employment of: (i) sub-wavelength photonic crystal structures, designating a roadmap for neuromorphic sub-wavelength photonics, (ii) plasmonic or plasmo-photonic structures, introducing the area of neuromorphic plasmonics or plasmo-photonics, and (iii) conventional silicon photonic circuitry, extending along the line of already defined neuromorphic silicon photonics. To analyze the energy and footprint efficiency perspectives of these photonic neuron technology roadmaps, the neuron functionalities have been first selected to comply with available solutions from the selected technology platform, using respective photonic circuits only when they bring significant performance advantages compared to the solution offered by the photonic crystal or plasmonic platform perspective. These three deployment paths are analyzed by assessing the minimum PD sensitivity and the total loss budget of the neuron in order to calculate the necessary laser output power, with the latter dictating the set of laser technologies that can form viable optical power supply options for each scenario. The footprint efficiency of the photonic neuron is then calculated by taking into account the physical dimensions for each selected device on top of the photonic substrate, assuming four silicon waveguide sections of 5 μm each for interconnecting the functional devices along the axon and

### Table III

<table>
<thead>
<tr>
<th>Type</th>
<th>IL</th>
<th>Power Cons. / μm</th>
<th>Vb</th>
<th>Reconf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[73],[74]</td>
<td>6 dB/ cm</td>
<td>0-100 nW</td>
<td>2-3</td>
<td>YES</td>
</tr>
<tr>
<td>Barium Titanite (BTO)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[75]MOS III-V</td>
<td>&lt;0.5 dB / 500 μm</td>
<td>1 nW</td>
<td>1-2</td>
<td>YES</td>
</tr>
<tr>
<td>[76]Post-fabrication trimming</td>
<td>0 dB</td>
<td>0 W</td>
<td>-</td>
<td>NO</td>
</tr>
<tr>
<td>[37]TO Si MZI</td>
<td>1.1 dB</td>
<td>28 mW</td>
<td>2.8</td>
<td>YES</td>
</tr>
<tr>
<td>[77]GeSbTe</td>
<td>1 dB / μm</td>
<td>0 W</td>
<td>-</td>
<td>YES</td>
</tr>
</tbody>
</table>

### Table IV

<table>
<thead>
<tr>
<th>Type</th>
<th>P_{in}</th>
<th>Power Cons.</th>
<th>Area</th>
<th>Bit Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>[40] DBR</td>
<td>6 mW</td>
<td>63.2 mW*</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>[43] DBR</td>
<td>15 mW</td>
<td>192.4 mW*</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>[44] 1D PhC</td>
<td>95 μW</td>
<td>678.6 μW*</td>
<td>90 μm²</td>
<td>N.A.</td>
</tr>
<tr>
<td>[45] 2D PhC</td>
<td>10 μW</td>
<td>94 μW*</td>
<td>112 μm²</td>
<td>10 Gb/s</td>
</tr>
<tr>
<td>[78]1D PhC</td>
<td>24 pW</td>
<td>4800 pW*</td>
<td>5.4 μm²</td>
<td>N.A.</td>
</tr>
<tr>
<td>[79]DFB</td>
<td>0.43 mW</td>
<td>4.2054 mW</td>
<td>L: 80 μm</td>
<td>25 Gb/s</td>
</tr>
<tr>
<td>[41] DFB</td>
<td>354 μW</td>
<td>34.3 mW</td>
<td>N.A.</td>
<td>80 Gb/s</td>
</tr>
<tr>
<td>[41] DFB</td>
<td>22 mW</td>
<td>308 mW*</td>
<td>L: 700 μm</td>
<td>N.A.</td>
</tr>
<tr>
<td>[81] DFB</td>
<td>22 mW</td>
<td>158 mW</td>
<td>L: 1000 μm</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

*Estimated values based on the reported wall-plug efficiencies and LIV curves.
Neuromorphic sub-wavelength photonics: In this case a linear neuron could almost entirely be fabricated using III-V PhC active devices supported by BTO technology for gating silicon photonic rings. That could in principle realize on chip weighting functions. A PhC based LEAP laser with a wall-plug-efficiency of ~11% [45] has been selected, with the modulation stage being offered by an EAM PhC based modulator with modulation speeds in the range of 40-56 Gbps and insertion loss of 3 dB [70]. For the PD section, we selected a bias- and amplifier-free nanophotonic PhC based PD with light-to-voltage conversion capabilities supporting data rates up to 40 Gb/s [59]. By assuming a safety optical loss margin of 6.5 dB to the losses of the modulator and weighting elements in order to account for possible additional losses stemming from coupling interfaces, weighting MRRs when loaded with non-volatile BTO gates [73], etc., a total loss budget of 10 dB is obtained. This means that a laser with a minimum output of power of 10 μW is needed in order to ensure a 1 μW of 40 Gb/s modulated optical power signal reaching the considered PhC PD. The total power consumption of this neuron is estimated to be 182.5 μW, which translates to 4.56 fJ/MAC efficacy. The total occupied area of the assessed neuron is estimated to be 7775 μm², leading to a footprint efficiency of 5.12 TMAC/s/mm².

Neuromorphic Plasmonics or Plasmo-photonic: This roadmap builds on the high-speed and ultra-small footprint of plasmonic modulator technology, assuming again weighting elements based on BTO gates and an additional loss safety margin of 6.5 dB. Opting for a SiGe PD at 100 Gb/s with a length of 20 μm [52], the total insertion loss that the linear neuron has to tolerate is 16.5 and 9.5 dB when a Plasmonic Organic Hybrid (POH) MZI [63] and a plasmonic RR [64] are employed for modulation purposes, respectively. This suggests that the required laser output power is 22 [81] and 3.5 mW [40] for the POH MZI and RR-based solutions, respectively, yielding respective energy efficiency values of 1.58 and 0.413 pJ/MAC. Considering the area requirements of POH MZI and RR devices, the footprint efficiency for both configurations ranges from 3 to 3.2 TMAC/s/mm².

Neuromorphic Silicon Photonics: Given that plasmonic and photonic crystal structures comprise certainly a longer-term technology perspective for light-enabled neurons, a technology roadmap relying entirely on silicon devices can offer a shorter-term viable perspective that could be eventually offered even by current CMOS and Multi-Project Wafer (MPW) fabrication platforms. Considering a SiGe PD at 56 Gb/s [53] a silicon MRR modulator with respective modulation bandwidth [72], a compact TO MZI [38] for delivering weighting functionality and 6.5 dB additional losses as a safety margin, the required optical input power ramps up to 18 mW suggesting an electrical laser power supply of 81 mW. In this case, the footprint and energy efficiency values are 1.45 pJ/MAC and 2 TMAC/s/mm², respectively.

Figure 9 provides an illustrative representation of the energy and footprint efficiency perspectives of the alternative single-layer neuromorphic photonic technology roadmaps, benchmarked against available analog and digital electronic neural network engines. Figure 9(a) shows various technologies scattered on a footprint efficiency ($\eta_F$) vs. energy-per-MAC ($E_{MAC}$) map. The map spans from MMAC/s/mm² to PMAC/s/mm² footprint efficiencies, covering 9 orders of magnitude, while the energy-per-MAC ranges from nJ/MAC to aJ/MAC, again extending along 9 orders of magnitude. Analog electronic representatives TrueNorth [9], [10], Neurogrid [13], [16] and HiCANN [17], [18], [19], reside in the lower left corner of the map, revealing a maximum $\eta_F$ of 0.3 GMACs/mm² and a minimum $E_{MAC}$ of 0.27 pJ/MAC, albeit, not for the same device, revealing a clear trade-off between the two metric parameters. Digital...
electronics, represented by NVIDIA, Google’s TPU, Graphcore and Groq, [82], [83], [84], [85], [86], shows an improvement over analog mostly in terms of $\eta$ by several orders of magnitude (3 to 6, depending on the specific comparison), with their $E_{\text{MAC}}$ performance remaining, however, in similar sub-pJ/MAC scales. The neuromorphic silicon photonic and plasmo-photonic solutions are shown to offer almost an additional one order of magnitude improvement in terms of $\eta$, going beyond TMAC/s/mm², even with on-chip lasers sources taken into account as real-estate consuming circuits. However, their energy efficiency remains in the sub-pJ/MAC range. An improvement in terms of energy can be obtained by transitioning to sub-wavelength photonics through III-V PhC structures, improving $E_{\text{MAC}}$ by 2-3 orders of magnitude and dropping to fJ/MAC ranges. Finally, a highly challenging target with $E_{\text{MAC}} = 10$ aJ/MAC for $\eta = 5$ PMACs/s/mm² that is close to the efficiency metrics projected in [25] is also illustrated in the far upper-right corner of the map in order to highlight the differences between the results of our analysis and the almost ideal technology scenario assumed in [25], helping at the same time to identify the technology advances that have to be materialized for elevating current technology standards to these performance levels. Migrating from current PhC structure technology to this staggering neuromorphic engine has to proceed along an energy reduction of close to 500x by either decreasing optical losses and/or increasing the laser wall-plug efficiency, accompanied by a reduction of three orders of magnitude in physical circuit and neuron dimensions for getting to a 5 PMACs/s/mm² footprint efficiency.

Figure 9(b) offers another perspective in benchmarking the different technologies by introducing a new figure of merit that merges both footprint and energy efficiency into a single quantity defined as $\eta E_{\text{MAC}}$. A significant improvement in $\eta E_{\text{MAC}}$ ratio is visible when transitioning from analog to digital electronics, with silicon photonics and PH MZI-based plasmo-photronics being at least comparable with digital engines. Migrating to PH MZI-based plasmo-photronics offers one order of magnitude better performance compared to the best-in-class Groq platform, which grows to up to 3 orders of magnitude improvement when InP photonic crystal structures are utilized in a neuromorphic sub-wavelength photonic hardware. This designates that even current light-enabled technology capabilities are suitable for boosting neuromorphic computational capacities, provided that a co-integration roadmap for bringing the individual circuits onto a common fabrication platform will be established. On top of these co-integration developments, the perspective of reaching breakthrough performance via six orders of magnitude improvements in the $\eta E_{\text{MAC}}$ ratio expected by a 10 aJ/MAC and 5 PMACs/s/mm² engine has to additionally evolve along respective breakthroughs in high-density integration and energy consumption requirements of constituent photonic crystal elements and corresponding fabrication platforms.

VI. CONCLUSIONS

In conclusion, we presented a comprehensive power consumption analysis of photonic neurons elaborating on their potential to lead to high energy- and footprint efficient neuromorphic engines. We formulated a mathematical framework for calculating neuron’s energy and footprint efficiency that has been successfully applied in demystifying the interdependence of different important circuitry specifications in the overall neuron performance and in identifying certain performance plateaus. Our theoretical framework reveals that energy efficiency optimization requires an optimal optical bandwidth value, stemming from the closely related parameters of neuron insertion losses and PD sensitivity requirements. Moreover, we clarify the important role of laser energy consumption when technologically feasible neuron fan-in values are utilized, considering also on-chip laser as a neuron real-estate consuming element. With the aim to establish a realistic performance framework for state-of-the-art technology capabilities, we provide a systematic review of best-in-class technology modules for lasing, modulating, weighting and receiving circuitry, subsequently attempting to map the energy and footprint efficiency toolkit onto viable hardware platforms. This analysis reveals the competitive benefits of current silicon photonic and neuromorphic plasmo-photonic technologies over today’s digital neural engines, highlighting that fJ/MAC energy efficiencies and TMACs/s/mm² compute densities have to evolve along InP-on-Si photonic crystal structures even with current device capabilities. This shapes a clear direction towards co-integrating the state-of-the-art structures into common yet high-performance fabrication platforms for immediately unleashing the expected energy and area savings to neuromorphic engines, with future research being required to yield orders of magnitude reductions in device size and power consumption for approaching the aJ/MAC and PMAC/s/mm² computational limits.

REFERENCES

Angelina R. Totović was born in Kragujevac, Serbia, in 1988. She received her BSc, MSc and PhD degrees in electrical engineering at the School of Electrical Engineering, University of Belgrade, Belgrade, Serbia, in 2011, 2013 and 2018, respectively.

From 2012 to 2018 she was a Teaching and Research Assistant at the School of Electrical Engineering, University of Belgrade. She joined WinPhos research group at the Aristotle University of Thessaloniki, Greece in 2018, where she currently holds the position of a Senior Researcher. She has authored/coauthored one chapter and more than 20 peer-reviewed research papers, and has been a reviewer for numerous IEEE and OSA publications. Her current research interests include neuromorphic photonics and exploitation of SOAs in various levels of photonic systems.

Dr. Totović is a recipient of several scholarships and awards for her scientific work. During her PhD studies, in 2016, she was awarded the Scholarship by Coimbra Group Universities through which she was a Visiting Researcher at the Photonics and Electromagnetics Group of the Department of Information Engineering at the University of Padova, Italy. In 2017 she was awarded the Optics and Photonics Education Scholarship by SPIE Society.

George Dabos was born in Thessaloniki, Greece, in 1989. He received his B.S degree from the Department of Informatics at the Aristotle University of Thessaloniki in 2012, and his M.S degree on “Networks - Communications and Systems Architecture” from the same department in 2014. He completed his PhD in “Integrated Photonic and Plasmonic Circuits for Optical Interconnect and Sensing Applications” in 2018 at the Aristotle University of Thessaloniki. So far, he authored and co-authored more than 40 peer-reviewed scientific articles published in IEEE, OSA and Nature as well as one patent. Dr. Dabos received an IAPP Marie Currie fellowship in 2015 working on the development of industry-driven PDA tools for Phoenix BV in the Netherlands (now with Synopsys). Currently, he is pursuing post-doctoral research in the field of neuromorphic photonics in WinPhos Reseach Lab. Dr. Dabos is currently working also as freelancer providing technical consulting service to Mellanox Technologies. His research interests focus on integrated photonic circuits and emerging technologies for datacom and neuromorphic computing.

Nikolaos Passalis is a postdoctoral researcher at the Department of Informatics, Aristotle University of Thessaloniki, Greece. He received the B.Sc. in Informatics in 2013, the M.Sc. in Information Systems in 2015 and the Ph.D. degree in Informatics in 2018, from the Aristotle University of Thessaloniki, Greece. He has (co-)authored 22 journal papers, 32 papers in
international conferences and contributed one chapter to one edited book. His research interests include deep learning, information retrieval and computational intelligence.

Anastasios Tefas received the B.Sc. in informatics in 1997 and the Ph.D. degree in informatics in 2002, both from the Aristotle University of Thessaloniki, Greece. Since 2017 he has been an Associate Professor at the Department of Informatics, Aristotle University of Thessaloniki. From 2008 to 2017, he was a Lecturer, Assistant Professor at the same University. From 2006 to 2008, he was an Assistant Professor at the Department of Information Management, Technological Institute of Kavala. From 2003 to 2004, he was a temporary lecturer in the Department of Informatics, University of Thessaloniki. From 1997 to 2002, he was a researcher and teaching assistant in the Department of Informatics, University of Thessaloniki. Dr. Tefas participated in 18 research projects financed by national and European funds. He is Area Editor in Signal Processing: Image Communications journal. He has co-authored 102 journal papers, 225 papers in international conferences and contributed 8 chapters to edited books in his area of expertise. Over 5400 citations have been recorded to his publications and his H-index is 36 according to Google scholar. His current research interests include computational intelligence, deep learning, pattern recognition, machine learning, digital signal and image analysis and retrieval, computer vision and robotics.

Nikos Pleros joined the faculty of the Department of Informatics, Aristotle University of Thessaloniki, Greece, in September 2007, where he is currently serving as an Associate Professor. He obtained the Diploma and the PhD Degree in Electrical & Computer Engineering from the National Technical University of Athens (NTUA) in 2000 and 2004, respectively. His research interests include optical interconnect technologies and architectures, neuromorphic photonics, photonic and plasmonic integrated circuits, optical technologies for disaggregated data center architectures and high-performance computing, as well as optical signal processing and optical switching. He has more than 300 archival journal publications and conference presentations including several invited contributions, while his work has been cited >3000 times (h-index=29, GS source). He has held positions of responsibility at several major conference committees including ECOC, OFC and SPIE Photonics West. Dr. Pleros has coordinated >7 FP7 and H2020 European projects having participated as partner in more than 10 additional projects. He has received the 2003 IEEE Photonics Society Graduate Student Fellowship granted to 12 PhD candidates world-wide in the field of photonics, while he was proud to (co-) supervise three more Fellowship winners during their PhD. Dr. Pleros was recently awarded the AUTH Excellence Award for his research project funding ID between 2016-2018.